Evaluation of Second-order Visual Features for Land-Use Classification

Romain Negrel, David Picard, Philippe-Henri Gosselin

Evaluation of Second-order Visual Features for Land-Use Classification

Romain Negrel*, David Picard* and Philippe-Henri Gosselin†
*ETIS/ENSEA - UCP - CNRS, F95014 Cergy, France
Email: romain.negrel,picard,gosselin@ensea.fr
†Texmex team, Inria, F35042 Rennes, France
Email: philippe.gosselin@inria.fr

***Abstract*—This paper investigates the use of recent visual features based on second-order statistics, as well as new processing techniques to improve the quality of features. More specifically, we present and evaluate Fisher Vectors (FV), Vectors of Locally Aggregated Descriptors (VLAD), and Vectors of Locally Aggregated Tensors (VLAT). These techniques are combined with several normalization techniques, such as power law normalization and orthogonalisation/whitening of descriptor spaces. Results on the UC Merced land use dataset shows the relevance of these new methods for land-use classification, as well as a significant improvement over Bag-of-Words.**

I. Introduction

In land-use classification of high-resolution overhead imagery, the most popular pipeline is composed of two parts: the extraction of image descriptors; and the use of statistical learning tools. To extract the descriptors, two types of methods are used: global descriptors; aggregated descriptors. The global descriptors commonly used are color descriptors [1] (*e.g.* histogram of RGB, HSV, CIE Lab) and homogeneous texture descriptors [1] (Gabor filter responses). The aggregated image descriptors are computed in two steps: the extraction of a set of local descriptors; and their aggregation in a single descriptor. The most polular local descriptors is the well known SIFT descriptor [2]. Many aggregation methods have been proposed and evaluated [1], [3–7]. Yang *et al.* [1], [3] evaluated numerous methods used in computer vision: the bag-of-visual-words [8] and its spacial extentions (Spatial Pyramid Match Kernel [9] and Spatial Co-occurrence Kernel [10]). Risojevic *et al.* [4] propose to use the cross-cerrelations between Gabor wavelet coefficient and quaternion framework for the representation of color images to compute the image descriptor. Recently, Cheriyadat *et al.* [5] propose to use a coding/pooling method [11] to aggregate the local descriptors into a single descriptor.

In this paper, we propose to evaluate recent methods of aggregated descriptors for visual representation of land-use classification in high-resolution overhead imagery [1]. This includes improved Fisher Vectors (FV) [12], Vectors of Locally Aggregated Descriptors (VLAD) [13], Vectors of Locally Aggregated Tensors (VLAT) [14], and many improvements that were recently proposed [15]. These methods are the extension of visual dictionaries approaches introduced by Bag-of-Words (BoW) [8], thanks to several key ideas. A first key idea is the introduction of the deviation approach, which was first motivated by statistical models from Fisher Kernels [12]. The idea is to consider the deviation between the local image model and a global model, rather than only the local image model. As a result, specific properties of an image are better emphasised. A second key idea is the introduction of second-order statistics, for instance using covariance data in addition to mean data. While the feature size is then significantly increased, recent techniques for high dimensionality reduction have solved this problem [13], [15]. Another key idea is the introduction of normalisation processing at the different levels of the tool-chain, like the power law [12] or cluster-wise component analysis [15]. A last key idea is to only consider features vectors compared with a linear similarity (*e.g.* dot product). These constraints allow the use of very efficient retrieval and learning techniques, like Stochastic Gradient Descent SVM [16]. Furthermore, in most cases this leads to methods with computational and memory complexity linear with the size of datasets, rather than quadratic complexity with non-linear features.

In this scope, we first propose a detailed presentation of these methods in Section II. Then, we present in Section III evaluation results on UC Merced land use dataset [1].

II. Image features

Most image features are obtained by a two steps scheme. The first step is to extract a set of local visual descriptors from the images. The most commonly used visual descriptors are highly discriminant local descriptors (HOG, SIFT, SURF, ...). Regions of interest can be selected by uniform sampling, or by automatic point of interest detection. The set extracted from an image is called a *bag*. We denote by $\mathbf{B}_i = \{\mathbf{b}_{ri}\}_r$ the set of descriptors $\mathbf{b}_{ri} \in \mathbb{R}^D$ in image i. The second step is to map the descriptors of the bag $\mathbf{B}_i$ into a single vector $\mathbf{x}_i \in \mathbb{R}^W$, known as the image feature.

A. Statistical Approaches

The first methods to map the descriptors in a feature are based on the statistical study of the distribution of descriptors in the bag. These approaches have been inspired by text retrieval methods. To study the distribution of descriptors, we use a visual codebook composed by C visual words. The visual codebook is generally computed by a clustering algorithm (*e.g.*, *k*-means) on a large sample of descriptors. A bag can then be described by a statistical analysis of occurrences of visual words.

The first method of this kind, named Bag of Words (BoW) [8] counts the number of descriptors belonging to

each cluster. The dimension of the feature is then C. Many extension of BoW has been proposed [11], for example in the classification of urban scenes in geo-referenced images [17]. These approaches obtain good results in similarity search and in images classification. However, to obtain good results with these methods it is necessary to use visual codebooks with very large dictionaries (about 100k visual word), and the use of a non-linear metric.

B. Model Deviation Approaches

Perronnin et al. [12] proposed a successful method called Fisher Vectors. This method uses a probability density function denoted by u_λ of parameters λ as model of the descriptors space. To describe the image, they compute the derivative of the log-likelihood of image descriptors to the model:

$$\mathcal{G}_\lambda^{\mathbf{B}_i} = \frac{1}{T}\nabla_\lambda \log u_\lambda(\mathbf{B}_i). \tag{1}$$

The authors propose to use a Gaussian Mixture Model (GMM) of parameters $\boldsymbol{\mu}_c$ and σ_c. Elements of the Fisher Vector for each Gaussian c can be written as:

$$\mathcal{G}_{\boldsymbol{\mu},c}^{\mathbf{B}_i} = \frac{1}{T\sqrt{\omega_c}}\sum_r \gamma_c(\mathbf{b}_{ri})\left(\frac{\mathbf{b}_{ri}-\boldsymbol{\mu}_c}{\sigma_c}\right), \tag{2}$$

$$\mathcal{G}_{\boldsymbol{\sigma},c}^{\mathbf{B}_i} = \frac{1}{T\sqrt{\omega_c}}\sum_r \gamma_c(\mathbf{b}_{ri})\left[\frac{(\mathbf{b}_{ri}-\boldsymbol{\mu}_c)^2}{\sigma_c^2}-1\right]. \tag{3}$$

Where $(\omega_c, \boldsymbol{\mu}_c, \sigma_c)$ are the weight, mean and standard deviation of Gaussian c, and $\gamma_c(\mathbf{b}_{ri})$ the normalized likelihood of $\mathbf{b}_{ri}$ to Gaussian c. The final feature is obtained by concatenation of $\mathcal{G}_{\boldsymbol{\mu},c}^{\mathbf{B}_i}$ and $\mathcal{G}_{\boldsymbol{\sigma},c}^{\mathbf{B}_i}$ for all Gaussians. Fisher Vectors achieve very good results [12]. However, Fisher Vectors are limited to the simple model of mixtures of Gaussians with diagonal covariance matrices. Moreover, the GMM algorithm is computationally very intensive.

Jegou *et al.* [13] proposed a simplified version of Fisher Vector by aggregating local descriptors, called Vectors of Locally Aggregated Descriptors (VLAD). They proposed to model the descriptors space by a small codebook obtained by clustering a large set of descriptors. The model is simply the sum of all centered descriptors $\mathbf{B}_{ci} = \{\mathbf{b}_{rci}\}_r \subseteq \mathbf{B}_i$ from image i and cluster c:

$$\boldsymbol{\nu}_{ci} = \sum_{\mathbf{b}_{rci}\in\mathbf{B}_{ci}} \mathbf{b}_{rci} - \boldsymbol{\mu}_c \tag{4}$$

with $\boldsymbol{\mu}_c$ the center of cluster c. The final feature is obtained by a concatenation of $\boldsymbol{\nu}_{ci}$ for all c. The feature size is $D \times C$.

Picard *et al.* [14] proposed an extension of VLAD by aggregating tensor products of local descriptors, called Vector of Locally Aggregated Tensors (VLAT). They proposed to use the covariance matrix of the descriptors of each cluster. Let us denote by "$\boldsymbol{\mu}_c$" the mean of cluster c and "$\mathcal{T}_c$" the covariance matrix of cluster c with $\mathbf{b}_{rci}$ descriptors belonging to cluster c:

$$\boldsymbol{\mu}_c = \frac{1}{|c|}\sum_i\sum_r \mathbf{b}_{rci} \tag{5}$$

$$\mathcal{T}_c = \frac{1}{|c|}\sum_i\sum_r (\mathbf{b}_{rci}-\boldsymbol{\mu}_c)(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)^\top, \tag{6}$$

with $|c|$ being the total number of descriptors in cluster c.

For each cluster c, the feature of image i is the sum of centered tensors of centered descriptors belonging to cluster c:

$$\mathcal{T}_{ic} = \sum_r (\mathbf{b}_{rci}-\boldsymbol{\mu}_c)(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)^\top - \mathcal{T}_c. \tag{7}$$

Each $\mathcal{T}_{ic}$ is flattened into a vector $\mathbf{v}_{ic}$. The VLAT feature $\mathbf{v}_i$ for image i consists of the concatenation of $\mathbf{v}_{ic}$ for all clusters:

$$\mathbf{v}_i = (\mathbf{v}_{i1}\dots\mathbf{v}_{iC}). \tag{8}$$

As the $\mathcal{T}_{ic}$ matrices are symmetric, only the diagonal and the upper part are kept while flattening $\mathcal{T}_{ic}$ into a vector $\mathbf{v}_{ic}$. The size of the feature is then $C \times \frac{D\times(D+1)}{2}$.

C. Normalization

Several normalization processing are proposed in the literature to enhance the quality of visual features. The most popular one is the power law normalization, which first raises each value to a power α, and then ℓ_2-normalize the resulting vector:

$$\mathbf{x}_i = \frac{\mathbf{v}'_i}{\|\mathbf{v}'_i\|}, \quad \forall j, \mathbf{v}'_i[j] = \text{sign}\left(\mathbf{v}_i[j]\right)\left|\mathbf{v}_i[j]\right|^\alpha, \tag{9}$$

with α typically set to 0,5. This normalization was first introduced for the final visual features [12], and more recently for the normalization of low-level descriptors, such as RootSIFT [18].

Another common improvement is the orthogonalization and/or whitening of vector spaces, using a Principal Component Analysis. This can be performed at different levels. For instance, this is a required pre-processing on low-level descriptors for Fisher Vectors [12]. It can also be used to normalize each cluster of the dictionary, as it is done for VLAD [19] and for VLAT [15].

In the case of VLAT, the processing is the following one. First, we compute the eigendecomposition of the covariance matrix of each cluster c:

$$\mathcal{T}_c = \mathbf{V}_c\mathbf{D}_c\mathbf{V}_c^\top, \tag{10}$$

where $\mathbf{D}_c$ is a real non-negative diagonal matrix (eigenvalues), and $\mathbf{V}_c$ is unitary (eigenvectors). Then we project the centered descriptors belonging to c on the eigenvectors:

$$\mathbf{b}'_{rci} = \mathbf{V}_c^\top(\mathbf{b}_{rci}-\boldsymbol{\mu}_c). \tag{11}$$

Combining eq.(11) and eq.(7), we get:

$$\begin{aligned}\mathcal{T}_{ic} &= \mathbf{V}_c^\top\left(\sum_r (\mathbf{b}_{rci}-\boldsymbol{\mu}_c)(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)^\top - \mathcal{T}_c\right)\mathbf{V}_c \\ &= \sum_r \mathbf{V}_c^\top\left((\mathbf{b}_{rci}-\boldsymbol{\mu}_c)(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)^\top\right)\mathbf{V}_c - \mathbf{D}_c \\ &= \sum_r \left(\mathbf{V}_c^\top(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)\right)\left(\mathbf{V}_c^\top(\mathbf{b}_{rci}-\boldsymbol{\mu}_c)\right)^\top - \mathbf{D}_c.\end{aligned}$$

The cluster-wise normalized VLAT feature of image i in cluster c is the sum of tensors of projected descriptors $\mathbf{b}'_{rci}$ belonging to cluster c, centered by $\mathbf{D}_c$:

$$\mathcal{T}_{ic} = \sum_r \mathbf{b}'_{rci}\mathbf{b}'^\top_{rci} - \mathbf{D}_c. \tag{12}$$

Fig. 1. UC Merced land use dataset.

III. Experiments

In this section, we present the result using FV, VLAD and VLAT features on UC Merced land use dataset [1].

A. Dataset

This dataset is composed of 256×256 pixels RGB images, with pixel resolution of one foot. They are manually classified into 21 classes, corresponding to various land cover and land use types: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Each class contains 100 images. Examples are presented in Fig. 1.

We evaluate methods using the same protocol as in [1]. This is a five-fold multi-classification protocol. We randomly split the dataset into five subsets, train using four subsets and test using the remaining one. The results presented in the following sections are then average classification accuracy over the five runs. We follow a one-versus-all classification strategy: for each class, we train a linear SVM classifier, and label each test image with the classifier that returns the highest classification score.

For all the following experiments, we ℓ_2 normalize low-level descriptors (no power law at this stage), and normalize final visual features using a power law with $\alpha = 0.5$.

B. Features comparison

We first evaluated the scale parameter of low-level descriptors. For this purpose, we used HOG descriptors at different scales: cells of 4 pixels, 6 pixels, 8 pixels and 10 pixels. Then, we carried out experiment with VLAD features for different dictionary sizes D, from 32 to 256 keywords. As presented in Table I, individual scales presents similar performance. Furthermore, this behaviour is the same with different dictionary size: in all cases, performance is improved by the size of the dictionary, but is similar from one scale to another. However, when combining all scales, the performance is improved. As a result, we always consider the combination of these four scales in the following experiments.

TABLE I. Classification accuracy (%) for various extraction scales of HOG descriptors and various visual codebooks size with VLAD feature.

D	32	64	128	256
4	84.5	86.7	88.5	89.0
6	84.8	86.0	87.6	89.0
8	83.2	86.7	87.7	88.6
10	83.8	86.5	87.8	88.7
4+6+8+10	**85.2**	**87.5**	**89.6**	**90.9**

TABLE II. Classification accuracy (%) for HOG and RGB descriptors and for various visual codebooks size with FV, VLAD, VLAT features.

		D			
		32	64	128	256
HOG	FV	88.3	90.0	91.2	91.8
	VLAD	85.2	87.5	89.6	90.9
	VLAT	**91.5**	**91.7**	**91.8**	**92.3**
RGB	FV	84.3	85.6	87.2	88.2
	VLAD	80.0	82.5	83.7	85.4
	VLAT	**87.5**	**88.1**	**88.1**	**88.9**

The second set of experiments compares FV, VLAD and VLAT using HOG descriptors and RGB descriptors [12], for different dictionary sizes D. Results are presented in Table II. If we compare HOG and RGB descriptors, HOG are more effective. Let note that we did not combine HOG and RGB descriptors because no normalization is performed on descriptor spaces in these experiments. Focusing on the dictionary size, improvement can be observed with VLAD, however, this is less significant for FV and VLAT. Finally, the best feature in these experiment is the VLAT feature.

TABLE IV. COMPARISON TO STATE-OF-THE-ART RESULTS.

Method	BOVW[1]	SCK[1]	BOVW+SCK[1]	ColorRGB[1]	ColorHLS[1]	Texture[1]	Our (VLAT)
Accuracy (%)	76.81	72.52	77.71	76.71	81.19	76.91	**94.3**
Method	SPCK[3]	SPCK+[3]	SPCK++[3]	Color O.F.[4]	Color O.D.[4]	Quater. O.D.[4]	Our (FV)
Accuracy (%)	73.14	76.05	77.38	81.10	84.86	85.48	93.8

TABLE III. CLASSIFICATION ACCURACY (%) FOR HOG AND RGB DESCRIPTORS AND FOR VARIOUS COMPRESSION RATIOS OF DESCRIPTORS WITH FV, VLAD, VLAT FEATURES ($D = 64$).

		d			
		128	96	64	32
HOG	FV	92.0	91.5	91.2	90.1
	VLAD	89.1	88.8	88.9	87.4
	VLAT	**92.5**	**92.7**	**92.7**	**91.8**
RGB	FV	-	89.3	89.4	87.2
	VLAD	-	86.5	86.0	84.6
	VLAT	-	**90.6**	**90.5**	**89.4**
HOG+RGB	FV	-	93.8	93.4	92.8
	VLAD	-	92.5	92.5	91.6
	VLAT	-	**94.1**	**94.3**	**93.8**

C. Descriptor spaces normalization

In this section, we analyse the normalization of descriptor spaces using methods based on Principal Components Analysis (PCA). The motivation is two-fold: improve the quality of features, but also combine several features from different descriptors. For FV, we considered different number d of component for the normalization and reduction of low-level descriptors. For VLAD and VLAT, this processing is performed for each cluster of the visual dictionary. Results are presented in Table III, were d is the number of principal component for the global PCA (FV) or for cluster-wise PCAs (VLAD and VLAT). Note that cases with 128 components are not presented for RGB descriptors, since the original descriptor only have 96 dimensions. Furthermore, the size of the dictionary is always set of 64.

Overall, if we compare to the previous set of experiments with $D = 64$, performance is increased in all cases, even if we reduce the size of descriptors to $d = 32$ dimensions. If we compare HOG and RGB descriptors, HOG are also more effective. Thanks to the normalization, we can create a relevant combination of HOG and RGB descriptors, which present the highest results, especially with VLAT features.

We present in Fig. 2 these results according to the size of features, using HOG and RGB descriptors, and dictionary sizes $D = 32$ and $D = 64$. If we focus on features with 1,000 dimensions, we can see that performance over 92% can be obtained with FV and VLAD methods. This can be compared to the results using BoW [1], which are from 77% to 82% with the same number of dimensions. Finally, we compare our results to the state of the art in Table IV. As we can see, model deviation approaches clearly outperform their methods by a fair margin of at least 9%.

IV. CONCLUSION

In this paper, we presented recent methods for computing visual features, as well as related normalization techniques. We carried out experiments to analyse the benefit of each new component on the UC Merced land use dataset [1]. Overall, all considered method (FV, VLAD and VLAT) showed high classification accuracies. Furthermore, a high accuracy can be obtain with small visual dictionaries (64 words), especially for FV and VLAT features. Descriptor space normalization using PCA-based methods improve the results, but also allows effective combination of different descriptors types (HOG and RGB in our experiments). Finally, when compared to BoW methods, these methods lead to large performance improvement, as it is seen for generalist image categorization. A new component that will be worth investigating in the future is the compression of visual features for large-scale indexing, and their evaluation on very large land-use datasets.

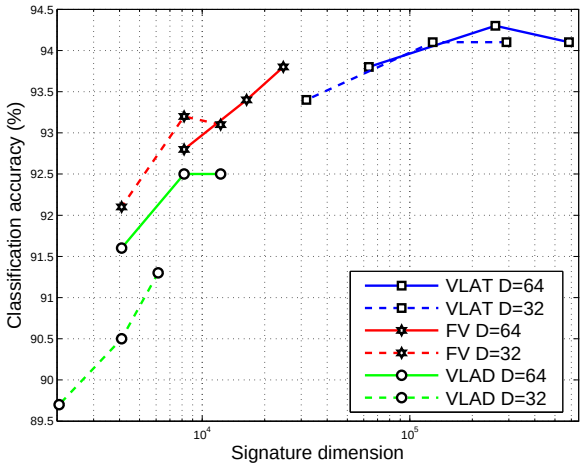

Fig. 2. Classification accuracy (%) as function of dimensionality reduction of descriptors (*i.e.*, $d \in \{96, 64, 32\}$) for various visual codebooks size and features.

REFERENCES

[1] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *ACM SIGSPATIAL GIS*, 2010.

[2] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. IEEE, 1999, pp. 1150–1157.

[3] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *ICCV*. IEEE, 2011, pp. 1465–1472.

[4] V. Risojevic and Z. Babic, "Orientation difference descriptor for aerial image classification," in *IWSSIP*. IEEE, 2012, pp. 150–153.

[5] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," pp. 439–451, 2014.

[6] Y. Zhang, X. Sun, H. Wang, and K. Fu, "High-resolution remote-sensing image classification via an approximate earth mover's distance-based bag-of-features model," pp. 1055–1059, 2013.

[7] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 3023–3034, May 2014.

[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, vol. 2, 2003, pp. 1470–1477.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[10] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010, pp. 3360–3367.

[12] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010, pp. 143–156.

[13] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. on PAMI*, vol. 1, pp. 3304–3311, 2012. [Online]. Available: http://hal.inria.fr/inria-00633013/en/

[14] D. Picard and P. Gosselin, "Efficient image signatures and similarities using tensor products of local descriptors," *CVIU*, vol. 117, p. 680687, 2013.

[15] R. Negrel, D. Picard, and P. Gosselin, "Web scale image retrieval using compact tensor aggregation of visual descriptors," *IEEE Multimedia*, vol. 20, no. 3, pp. 24–33, 2013.

[16] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *COMPSTAT*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Springer, August 2010, pp. 177–187. [Online]. Available: http://leon.bottou.org/papers/bottou-2010

[17] C. Iovan, D. Picard, N. Thome, and M. Cord, "Classification of Urban Scenes from Geo-referenced Images in Urban Street-View Context," in *ICMLA*, vol. 2, United States, 2012, pp. 339–344. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00794980

[18] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*. IEEE, 2012, pp. 2911–2918.

[19] J. Delhumeau, P.-H. Gosselin, H. Jegou, and P. Perez, "Revisiting the vlad image representation," in *ACM Multimedia*, Barcelona, Spain, October 2013.